Sentiment Analysis of COVID-19 Tweets by Text Classification

Fangyu Lo
University of California, San Diego
MS Business Analytics 21'
falo@ucsd.edu

Qing Gong
University of California, San Diego
MS Business Analytics 21'
qgong@ucsd.edu

Abstract

COVID-19 is a contagious disease caused by coronavirus 2 (SARS-CoV-2). It is originally known as Corona Virus of 2019 and has been declared as a pandemic by World Health Organization (WHO) on 11th March 2020. As the number of cases raised exponentially around the world, anxiety, stress, and other mental conditions also became a huge problem. This paper aims to analyze how sentiment of people changes during the pandemic time on social media. This research used several text classification and deep learning techniques and compared findings using different models. The research demonstrates that TensorFlow Keras model has the best performance out of four classification techniques, using this data set. Moreover, we found that negative tweets are strongly correlated with rising COVID cases within an area.

1 Introduction

COVID-19 is a new disease, caused by a novel (or new) coronavirus that has not previously been seen in humans. Its first case was identified in Wuhan, China in December 2019. Since then, many countries have been taken place in lockdowns and social distancing, and social media has become an important channel for people to express their emotions. It has been reported that COVID-19 and the social distancing became a source of depression, stress, and anxiety. With the lockdowns, social media is now the prime way for most people to get connected. Thus, the highest activity has been reported on many social media platforms.

Sentiment analysis has now become one of the most popular topic in the field of natural language processing and is also considered a powerful tool. As the social media nowadays are reporting higher than ever activities, Twitter has become a great source of information to gather social data. Individuals express their emotions on these platforms and discuss their opinions and perspectives towards these topics since social media allows people to share their opinions, views, and concerns.

This paper focuses on building, training, and finding the most accurate text classifier to analyze the sentiment of tweets. Moreover, this paper tries to combine how the trends of sentiment change over time with the trends of cases in each location. The paper is organized in the following structure: In section 2, the data used to analyze will be introduced. In section 3, the word embedding methodologies applied will be discussed. In section 4, the prediction models will be introduced and the performance will be evaluated. In section 5, the trends of sentiment changes over time at each location in the US will be presented. In section 6, further application of the text classification result introduced in section 4 will be compared with the sentiment trends. Finally, In section 7, we will conclude the research and summarize the paper.

2 Data Description

2.1 Data Overview

The data used to analyze in this paper is extracted from the following Kaggle webpage:

https://www.kaggle.com/datatattle/covid-19-nlp-text-classification

Twitter is a social networking and microblogging service that allows users to post real time messages, called tweets. Tweets are short messages, restricted to 140 characters in length. Our data contains tweets pulled from Twitter, and each tweet is labeled by a human annotator as positive, negative or neutral. The names and usernames have been given codes to avoid any privacy concerns. In the data, the columns include the `Location` (user location), `TweetAt` (tweet time), `OriginalTweet`

Sentiment	Tweet Count
Extremely Positive	6,073
Positive	12,369
Neutral	8,332
Negative	10,958
Extremely Negative	6,073

Table 1: Sentiment distribution.

(tweet context), and `Sentiment` (sentiment label). The data set is also split into 41,157 for training and 3,798 for testing purposes. The models were trained and tested on the text of `OriginalTweet` with the labels of `Sentiment`.

2.2 EDA

To reveal the underlying hidden patterns in the data, Exploratory data analysis was taken place before building models.

Combining train set and test set, there are 41,157 tweets in total. The majority of the tweets are positive tweets, including tweets labeled as 'positive' and 'extremely positive'. The distribution of sentiment of tweets are generally balanced. There are slightly more negative tweets in the test set than in the train set but they are generally similar.

We also checked for the most common words in the data set. Words such as `people`, `grocery` and `store` appears a lot, which is reasonable because these are the topics people care most about during the pandemic. There are also some interesting common words related to the pandemic such as `hand sanitizer`, `covid`, `panic`, and `help`.

Figure 1: Most Common Words

3 Word-Embedding Methodologies

The texts of the tweets were preprocessed before trained. The preprocessing mainly consists of tokenization, and stemming. The tweets were first removed URL links, HTML tags, digits, hashtags, and mention tags. Then, they were split into single words, removed punctuation and stopwords, as well as converted to word's stem. After preprocessed, the texts were trained using word embedding techniques, which include TF-IDF, Word2Vec, and GloVe model. Moreover, we also tried Keras Embedding Layer to analyze the data. Since the these word-embedding methods each have their pros and cons, we decided to use all of them to create vector representation for the words, build prediction models on all of them, and compare their respective performances.

TF-IDF stands for "Term Frequency — Inverse Document Frequency", mainly calculating how many times a word appears in a document and the inverse document frequency of the word across a set of documents. This technique helps quantify a word in documents. and computes the importance of a word in a document and corpus by giving weight to each word. The formula is outlined below:

```
TF-IDF = Term Frequency (TF) * Inverse
Document Frequency (IDF)
```

TF-IDF works by increasing proportionally to the number of times a word appears in a document but is offset by the number of documents that contain the word.

In this paper, we used the `TfidfVectorizer` in the Scikit-learn feature extraction package to calculate the TF-IDF for words in tweets. We ignored terms that had a document frequency strictly lower than 50. Also, We smoothed IDF weights by adding one to document frequencies. The shape of the features consisted of 1,909 columns.

Word2Vec is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets. It works by using Skip Gram and Common Bag Of Words (CBOW) methods. The two methods both involve Neural Networks. Skip Gram works well with a small amount of data and is found to represent rare words well. On the other hand, CBOW is faster and has better representations for more frequent words. Word2Vec is one of the most popular techniques to learn word embeddings.

In this paper, we used the `Word2Vec` model from the Gensim library. We first build the vocabularies by feeding training data. We pruned the internal dictionary by specifying the minimum word count to 50. We set the vector dimension to be 200.

GloVe stands for "Global Vectors". GloVe (Pennington et al., 2014) is an unsupervised learning algorithm for obtaining vector representations for words. The advantage of GloVe is that, unlike Word2vec, GloVe does not rely just on local statistics (local context information of words). It is a word vector technique that leverages both global and local statistics of a corpus to come up with a principled loss function which uses both these. The training of GloVe is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

In this paper, we applied the pre-trained Twitter word vectors with 27B tokens and 200 dimensional vectors downloaded from:

`https://nlp.stanford.edu/projects/glove/`

TensorFlow Keras is a Python deep learning API which offers an embedding layer that can be used for neural networks on text data. It is designed for large-scale training and prediction purposes. The Embedding layer is initialized with random weights and will learn an embedding for all of the words in the training dataset. The embedding layer is implemented in the form of a class in Keras and is used as a first layer in the sequential model for NLP tasks. We tested the hyper-meters such as the number of layers, number of nodes and picked the hyper-meters generating best accuracy manually.

4 Models and Predictions

We used the word vectors built by the TF-IDF, Word2Vec, and GloVe methods to train logistic regression models and predict the sentiments of the Tweets. We also trained the word vectors by a Tensorflow Keras neural network model. We will first introduce the four models we built and then discuss their performances.

4.1 Modeling

Two types of models were used to train the Tweets data, logistic regression, and neural network. In this section, the fundamental concepts of the two modeling techniques will be explained with the hyper parameters used in training the models. Moreover, the data contains five sentiment labels, including `Extreme Positive`, `Positive`, `Neutral`, `Negative`, and `Extreme Negative`. We first trained and tested the models on the five labels. After doing so, we combined `Negative` and `Extreme Negative` labels into one `Negative` label, as well as `Extreme Positive` and `Positive` into one `Positive` label. We then trained and tested the models on the three sentiment labels. Totally, we built eight models using both logistic regression and neural network.

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable (target) is binary or categorical. Like all regression analyses, logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between the dependent variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

In this paper, we trained three logistic regression models using the TF-IDF vector metrics, Word2Vec vector metrics, and GloVe vector metrics. We applied the `LogisticRegressionCV` function from the Scikit-learn library. The function applies **k-fold cross-validation with the logistic regression**. It works by first dividing the train data set into different train/validation set combinations and training them using logistic regression. The models across different splits of data will then be tested to bring the best model with the greatest performance.

We used 5-fold cross-validation with a list of C values and set the maximum iteration as 100,000,000. **The performance matrix we used to evaluate the model performance is accuracy rate**, which is the percentage of correct predictions for a given data set.

Neural Network is a network composed of artificial neurons or nodes. Like logistic regression we used earlier, Neural Network can also be used to perform the classification of data. Indeed, logistic regression is a special case of a neural network, which has no hidden layers. Logistic regression uses the sigmoid activation function, which relates the input to the output. Comparing to logistic regression, Neural Network is better in capturing more complex non-linear relationship, with respect to the features.

There are two ways of building models in Keras. We can use Sequential API or the Functional API. In our research, we chose to use the Sequential API to build the model. Sequential API is used to build models as a simple stack of layers. We first initiated the model and add layer one by one. Each layer of the stack of layers has exactly one input tensor and output tensor. For the hyper parameters, we chose to use EPOCHS equals 2, BATCH SIZE equals 32, embedding-dim equals 16 and units equals 256.

4.2 Performance Evaluation

We evaluated the model performance based on accuracy rate, macro F1-score, and micro F1-score. The performance metrics were summarized in Table 2. The neural network performed the best on both the five-label model and three-label model with accuracy rates of 0.75 and 0.85. The TF-IDF embedding method brought the greatest performance for the logistic regression, 0.63 accuracy rate for the five-label model and 0.8 accuracy rate for the three-label model. The logistic regression for the Word2Vec and GloVe models had similar performances within each of the five-label and three-label models.

Overall, we can see an increase in accuracy rate and F1-score when we moved from the five-label model to the three-label model. It is reasonable since the models would not need to distinguish between extremely negative tweets and negative tweets as well as extremely positive tweets and positive tweets.

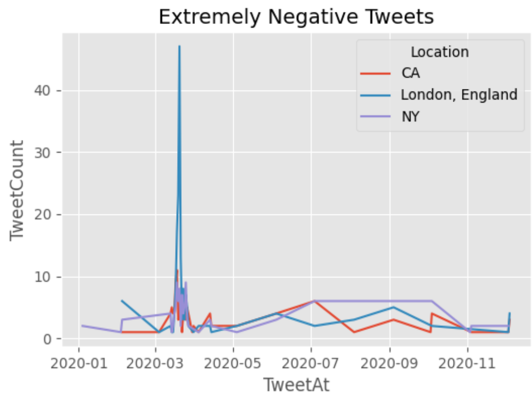
5 COVID-19 Sentiment Trends

We merged the train and test data and grouped the data based on the time users tweeted and the location of the tweeted users. The highest number of tweets in the US came from California (1,114 tweets) and New York (1058 tweets). Also, London had more than 1,000 tweets (1,266 tweets). We decided to compare the number of extremely negative tweets in the three locations because they are more re-presentable and because the extremely negative sentiment on social medial should be a warning sign that society is in panic or terror. Figure 2 is a line chart describing this extremely negative sentiment during the pandemic.

You can see that in March 2020 when the western world encountered the first wave of COVID-19 outbreak and several governments mandated stay-at-home order or city lockdown, the number of extremely negative tweets reached a record high. However, it did not last long. This chart is a good indicator of the fear and panic inside our community.

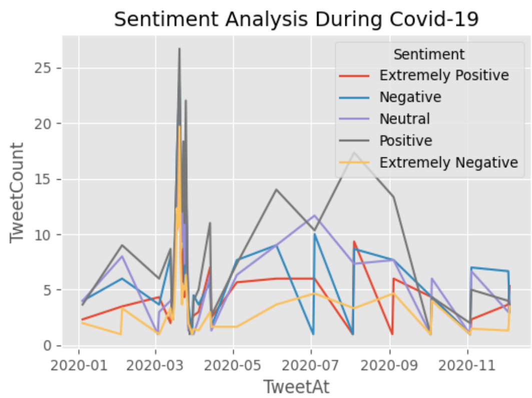
To investigate the trends of all sentiments at once,

Figure 2: Extremely negative tweet time trend.



we created a chart that summarized the five sentiments across all locations in the same COVID-19 period as Figure 3. From the charts, we can see that the sentiment of social media users fluctuated a lot during the pandemic. Sometimes positive sentiment followed immediately by extremely negative thoughts, but sometimes they jumped high together. The trends demonstrate that people are unsure about the future of the pandemic, and the two contrary emotions might affected by the news, variation of the stock markets, and the rise of infected cases.

Figure 3: Twitter sentiment trend during COVID-19.



6 Further Applications

In the previous section, we proved that the sentiment across social media is useful and effective in tracking the panic and fear in society. In this section, we will explain the possible application of the prediction models we built.

Using the 5-label TensorFlow Keras neural network models, we compared the predicted extremely negative tweets with the actual extremely negative

Five-Label Model	Accuracy	macro F1-score	micro F1-score
TF-IDF logistic regression	0.63	0.64	0.63
Word2Vec logistic regression	0.42	0.42	0.42
GloVe logistic regression	0.44	0.43	0.44
Keras neural network	0.75	0.76	0.75
Three-Label Model	Accuracy	macro F1-score	micro F1-score
TF-IDF logistic regression	0.80	0.78	0.80
Word2Vec logistic regression	0.65	0.61	0.65
GloVe logistic regression	0.65	0.61	0.65
Keras neural network	0.85	0.84	0.85

Table 2: Model performance evaluation metrics.

Figure 4: Predicted extremely negative tweets by 5-label model

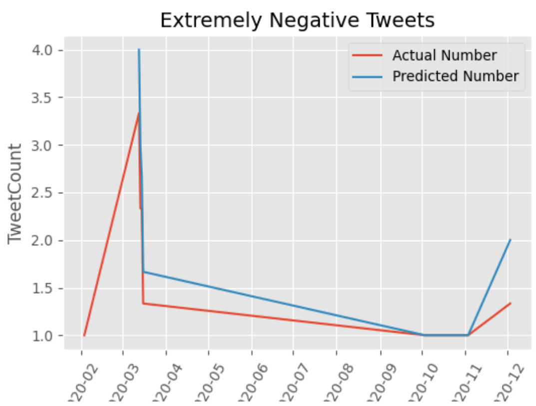
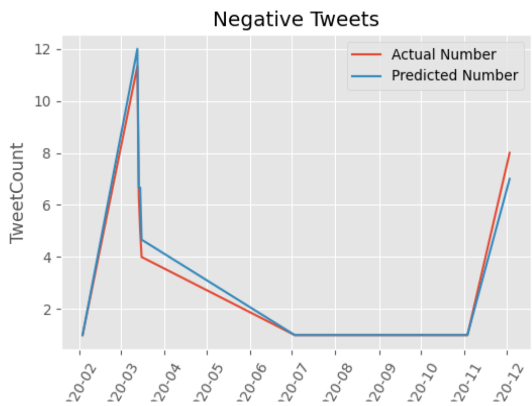


Figure 5: Predicted negative tweets by 3-label model



tweets as in the Figure 4. You can see the models helps predict panic, fear, and negative thoughts. For the government department, tracking the negative sentiment or the fluctuation of the sentiment trend is very important as it helps us maintain social security, the stability of the economy, and people's mental health. We also see some recent cases that extremists gathered in social media groups for planning national terrorist attacks. Thus, if our prediction could be extended to real-world use, it is expected to prevent the terrorist and heal further panic in the society.

We also used the 3-label TensorFlow Keras neural network to create a chart compare the actual negative tweets and the predicted negative tweets as Figure 5. You can see from the line charts that this model predicted the negative sentiment very well. It could also be applied to real-world cases to raise the awareness of the depression in the community.

7 Conclusion

We presented our results for sentiment analysis on Twitter data during COVID-19 pandemic and combined the findings with the COVID cases numbers and policies taken place during that time. We used logistic regression model using the TF-IDF vector metrics, Word2Vec vector metrics, and GloVe vector metrics. For five-label models, these models respectively reports accuracy score of 0.63, 0.42 and 0.44. We investigated a Neural Network model: Tensorflow Keras and it reports an accuracy score of 0.75. For three-label models, TF-IDF, Word2Vec, Glove and Keras respectively reports accuracy score of 0.80, 0.65, 0.65 and 0.85. TensorFlow Keras model has the best accuracy score for both five-label models and three-label models, which corresponds to our assumption in section four.

Combining these sentiment analyses with real-world cases in different locations, we generalized that negative tweets are strongly correlated with lockdown policies and rising COVID cases number. These findings reasonably reflected that people are unsure about their lives during this pandemic and emotions might be affected by news of the pandemic, including new cases, fluctuations in the stock market, etc.

In future work, we will explore even more on richer linguistic analysis such as better lemmatization and cleaning. More than that, our work could be possibly extended to other areas in real-world scenarios. —

References

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

—